Proc. Natl. Acad. Sci. USA
Vol. 74, No. 10, pp. 4126–4129, October 1977
Chemistry

Local interactions in bends of proteins

(protein x-ray structures/conformational energy calculations/dipeptides/long-range interactions)

S. SCOTT ZIMMERMAN* AND HAROLD A. SCHERAGA

Department of Chemistry, Cornell University, Ithaca, New York 14853

Contributed by Harold A. Scheraga, July 20, 1977

ABSTRACT **Calculated probabilities of bend formation in 47 amino acid sequences of *N*-acetyl-*N'*-methylamide dipeptides, determined from a statistical mechanical analysis using empirical conformational energies, were compared with the observed fraction of bends formed in the same 47 dipeptide sequences in the x-ray structures of 20 globular proteins. Agreement between the calculated and observed fraction of bends was found for 26 dipeptides, suggesting that, for those particular dipeptide sequences, local interactions dominate over long-range interactions in determining conformational preference. Seven dipeptide sequences, all of which contained a Gly residue, had a significantly higher calculated than observed bend preference, indicating the strong influence of long-range and/or solvent interactions in those sequences. Of the 14 sequences for which the calculated was significantly less than the observed bend fraction, 13 dipeptide sequences contained at least one polar residue (Ser, Asn, or Asp) and/or an aromatic residue (Phe or Tyr), suggesting that solvent effects may play an important role in dictating the conformation in these sequences. The analysis of dipeptide sequences in the twenty globular proteins also indicated that the 4 → 1 hydrogen bond is not a dominant factor in stabilizing bends in proteins, and that most dipeptide sequences are capable of forming several types of bend conformations.**

Chain reversals (also called β bends, β turns, etc.) have been shown (1–8) to be important conformations in the three-dimensional structure of peptides and proteins. Much theoretical work has been done in an effort to understand the factors that stabilize bends (1, 3–17), and algorithms have been developed to try to predict the locations of bends in the amino acid sequences of proteins (3, 18–21). In earlier studies, evidence has been presented (2, 3, 5, 7–12, 14, 15) which indicates that bends can be stabilized by local interactions, i.e., those within the dipeptide sequence of the $i+1$ and $i+2$ residues of the bend (15), but that medium- and long-range interactions may be important for the stability of some bends (13, 22).

In this paper, we analyze the occurrence of bends among dipeptide sequences in the x-ray structures of twenty globular proteins. The fraction of a particular dipeptide sequence that occurs in bends in proteins is compared with the bend probability for the same dipeptide sequence calculated in earlier studies (16, 17) from statistical mechanics and empirical conformational energies. Because the conformational energies were determined (15–17) for *isolated* *N*-acetyl-*N'*-methylamide dipeptides, the calculated bend probabilities do not include the long-range and solvent interactions present in globular proteins. Thus, it is expected that information on the similarities and differences between the calculated and observed bend fractions will provide information about the relative contributions of solvent, short-range, and long-range interactions in stabilizing bend conformations in globular proteins. Furthermore, an analysis of dipeptide sequences in proteins may provide a better understanding of the properties of bends, which may help in the prediction of their occurrences in proteins of unknown structure.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

METHODS

Nomenclature and conventions are those adopted by an IUPAC-IUB Commission (23). All non-Gly residues are taken in the L configuration.

A bend is defined (15, 16) as a conformation in which the distance between the C^α of a residue i and the C^α of a residue $i+3$ along the peptide chain is ≤ 7 Å (3). If both residues $i+1$ and $i+2$ are part of an α-helix, the dipeptide segment is not considered a bend even though the $C^\alpha_i \cdots C^\alpha_{i+3}$ distance is ≤ 7 Å. Making the assumptions that bond lengths and bond angles are relatively constant in the peptide groups ($C^\alpha HC'ONHC^\alpha H$) and that the peptide groups are approximately planar, the $C^\alpha_i \cdots C^\alpha_{i+3}$ distance is a function only of the backbone dihedral angles ϕ and ψ of the dipeptide segment made up of residues $i+1$ and $i+2$. In the *N*-acetyl-*N'*-methylamide dipeptides, a bend is defined as a conformation in which the distance between the terminal CH_3 groups is ≤ 7 Å.

The definitions of an α-helix and of extended structure in proteins are given elsewhere (15). The analysis of the protein data excluded dipeptide sequences in which both residues were in an α-helix, but included extended structures (15) unless otherwise noted.

Calculated Bend Probabilities. Bend probabilities $P_{Q,b}$ and $P_{Z,b}$ are those calculated in earlier studies (15–17). Here we briefly describe the methods and definitions used for those calculations; further details are available in the original papers (15–17). The total conformational energy E_i of the ith energy minimum was calculated using ECEPP [empirical conformational energy program for peptides (24)] and energy minimization procedures (25). Starting points for energy minimization were combinations of all single-residue minima (26) and bend conformations (16). Bend probabilities are defined by the equations

$$P_{Q,b} = (1/Q) \sum_{i=1}^{l} \exp[-\Delta E_i/RT] \quad [1]$$

and

$$P_{Z,b} = (1/Z)(2\pi RT)^{k/2} \sum_{i=1}^{l} (\det \mathbf{F}_i)^{-1/2} \exp[-\Delta E_i/RT] \quad [2]$$

in which ΔE_i is the conformational energy at the ith minimum (relative to the global minimum taken as zero) and the summations are taken over all low-energy bend conformations for the particular molecule, in which

$$Q = \sum_{i=1}^{n} \exp[-\Delta E_i/RT] \quad [3]$$

* Present address: Department of Chemistry, University of Arkansas at Little Rock, 33rd and University Ave., Little Rock, AR 72204.

and

$$Z = (2\pi RT)^{k/2} \sum_{i=1}^{n} (\det \mathbf{F}_i)^{-1/2} \exp[-\Delta E_i/RT] \quad [4]$$

and in which l is the number of low-energy bend conformations, n is the total number of low-energy minima (i.e., those within 3 kcal/mol of the global minimum), k is the number of variable dihedral angles, $\mathbf{F}_i$ is the matrix of second derivatives at the ith minimum, R is the gas constant, and T is the absolute temperature, taken as 300 K in these calculations. Eqs. 1 and 2 are two approximations of classical statistical mechanical methods for determining probabilities, in which the Q-approximation (Eq. 1) involves only energies at each local minimum, while the Z-approximation (Eq. 2) involves both energy and librational entropy at the local minima. The conditions of validity of these approximations were discussed earlier (15).

Conformational energy analyses have been carried out on 54 *N*-acetyl-*N'*-methylamide dipeptides (blocked dipeptides) (16, 17). The blocked dipeptides included those of the types Pro-X and X-Pro (16), Ala-X and X-Ala (17), Gly-X and X-Gly (17), and Ser-X and X-Ser (17), for which, in each set, X *usually* included, but was not limited to, the residues Ala, Asn, Asp, Gly, Phe, Pro, Ser, Tyr, and Val. Solvent was not included in the calculations.

Observed Bend Fractions. The x-ray structures of 20 globular proteins were examined to determine the conformational preferences of each of the dipeptide segments corresponding to the sequences of the same 54 blocked dipeptides for which calculations have been performed (16, 17). The 20 protein x-ray structures are listed elsewhere (15). The observed, or empirical, bend probability (fraction) $P_{E,b}$ is defined as (15)

$$P_{E,b} = N_b/N \quad [5]$$

in which N_b is the number of occurrences of bends and N is the total number of occurrences of a particular dipeptide sequence among the 20 proteins, excluding those dipeptide segments in which both residues are part of an α-helix. Helical structures were excluded because of specific, observable long-range interactions (15), i.e., i+4 to i hydrogen bonds. Extended structures were not excluded from the analysis because, in many cases, no specific long-range interactions can be observed (15).

Because the values of $P_{E,b}$ are to be compared with $P_{Q,b}$ and $P_{Z,b}$, the statistical error in $P_{E,b}$ must be estimated. For this purpose, we use the Bayesian methods employed previously (27). Upper and lower limits in the value of $P_{E,b}$ are determined in the following manner. The logarithm of the odds for a dipeptide sequence to be in a bend can be estimated by $L = \ln[P_{E,b}/(1 - P_{E,b})]$, and the variance in this estimate is given by $\sigma^2 = (N_b + 1)^{-1} + (N - N_b + 1)^{-1}$. Therefore, the upper and lower limits of $P_{E,b}$, within two standard deviations (~95% confidence) are given by

$$P_{E,b}^{\text{upper}} = \{1 + \exp[L + 2\sigma]\}^{-1} \quad [6]$$

and

$$P_{E,b}^{\text{lower}} = \{1 + \exp[L - 2\sigma]\}^{-1} \quad [7]$$

For the purposes of discussion in this paper, if the value of $P_{Q,b}$ (or $P_{Z,b}$) is between $P_{E,b}^{\text{upper}}$ and $P_{E,b}^{\text{lower}}$, the calculated and observed bend fractions are considered to be in agreement.

It is of interest to examine bend conformations for the possible presence of a 4 → 1 hydrogen bond. Such a hydrogen bond is considered to exist if the distance $O_i \cdots N_{i+3}$ is ≤3.2 Å, in which O_i is the backbone carbonyl oxygen atom of residue i and N_{i+3} is the backbone amide nitrogen atom of residue i+3.

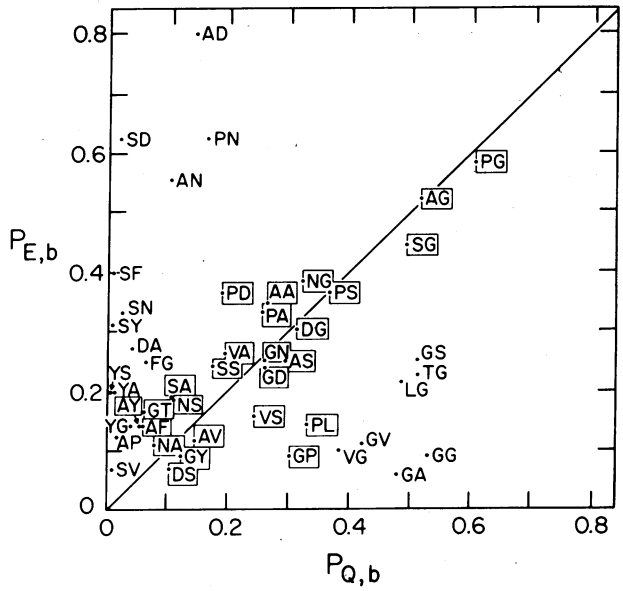

FIG. 1. A plot of the bend fraction $P_{E,b}$ for dipeptide segments, from an analysis of 20 protein x-ray structures, versus the bend probability $P_{Q,b}$ of *N*-acetyl-*N'*-methylamide dipeptides, calculated from an empirical conformational energy analysis. The single letter abbreviations of amino acid residues are (23) A = Ala, D = Asp, F = Phe, G = Gly, L = Leu, N = Asn, P = Pro, S = Ser, T = Thr, V = Val, and Y = Tyr. Dipeptides in which $P_{Q,b}$ agrees with $P_{E,b}$ within two standard deviations are framed by boxes.

RESULTS AND DISCUSSION

Dipeptide Sequences in Proteins. In the set of 20 protein x-ray structures, there are a total of 2654 dipeptide segments for which a $C^{\alpha}_i \cdots C^{\alpha}_{i+3}$ distance can be determined and in which one or both residues are *not* part of an α-helix. Of these 2654 dipeptides, 649 or 24% are in a bend conformation. Among these bends, 203 or 31% have a 4 → 1 hydrogen bond. These data indicate that bends are common conformational features of globular proteins, and that a 4 → 1 hydrogen bond is a common feature of bends. However, because a 4 → 1 hydrogen bond is present in only 31% of the bends, such a hydrogen bond is not an important factor is stabilizing most bends (8).

The occurrence of the individual bend types, defined in earlier papers (1, 3, 8, 16), is as follows: type I, 312; I', 12; II, 63; II', 24; III, 133; III', 15; IV, 39; V, 0; V', 0; VI, 5; VII, 46. In the globular proteins, the individual dipeptide sequences usually occur in more than one bend type. For example, the sequence Pro-Gly occurs 12 times, seven of which are bends, with type I occurring once, type II twice, and type III four times. Ala-Gly has a total of 23 occurrences, with 12 being in bends, and with type I occurring twice, type II seven times, type II' once, type III once, and type VII once. The important conclusion is that a particular bend type rarely is the only type to occur for a particular dipeptide sequence. The same conclusion was drawn from the results of conformational energy calculations; low-energy minima of a particular dipeptide usually include several bend types (16, 17). This implies that the prediction of a particular bend type at a particular location in a protein will require a knowledge of the specific long-range interactions that influence the dipeptide sequence.

Comparison of $P_{Q,b}$ with $P_{E,b}$. Fig. 1 shows a comparison of the calculated ($P_{Q,b}$) and observed ($P_{E,b}$) bend fractions for 47 dipeptides. Calculations of $P_{Q,b}$ were performed on seven other dipeptides, but those dipeptide sequences each occur only six or fewer times in the set of 20 protein x-ray structures, and

therefore were not included in the comparison because of the high statistical error.

The dipeptides for which the values of $P_{Q,b}$ and $P_{E,b}$ agree within two standard deviations are indicated by the boxes around the amino acid single letter abbreviations. As can be seen, 26 dipeptides are found to agree and 21 dipeptides to disagree in their values of $P_{Q,b}$ and $P_{E,b}$. Some representative values of $P_{E,b}^{upper}$ and $P_{E,b}^{lower}$ are 0.32 and 0.80, respectively, for Pro-Gly, for which $P_{E,b} = 0.58$ and $N = 12$; 0.33 and 0.71, respectively, for Ala-Gly, for which $P_{E,b} = 0.52$ and $N = 23$; 0.03 and 0.41, respectively, for Ala-Pro, for which $P_{E,b} = 0.13$ and $N = 8$.

If a similar comparison between $P_{E,b}$ and $P_{Q,b}$ is made by excluding extended structures as well as helical structures from the determination of $P_{E,b}$, 23 dipeptides agree in their values of $P_{Q,b}$ and $P_{E,b}$, 19 dipeptides disagree, and 12 dipeptides lack sufficient data for the comparison. The general conclusions which follow in the remainder of this paper are based on the results in which only helical structures were excluded. These same conclusions, however, result from the analysis with extended and helical structures excluded.

It is instructive to analyze the possible sources of the agreements and discrepancies between the bend fractions observed in proteins and those calculated for blocked dipeptides. Possible sources of the discrepancies include (*a*) deficiencies in the methods and/or parameters used in the conformational energy calculations, (*b*) approximations in the statistical mechanical analysis in calculating $P_{Q,b}$, (*c*) failure to include solvent in the calculations, (*d*) lack of long-range interactions in the calculations on blocked dipeptides (because long-range interactions are present in the globular proteins), and (*e*) experimental error in the x-ray crystallographic data. Explanations *a–d* are discussed in the following paragraphs. Explanation *e* will not be discussed because we have no way of assessing such errors.

(*a*) One known deficiency in the parameters, which may cause disagreement between $P_{E,b}$ and $P_{Q,b}$ in some dipeptides, involves the NH···OC hydrogen bond energy. This has been discussed in detail elsewhere (28). However, the generally good agreement (16, 17, 28) between the calculations on the blocked dipeptides and experimental observations of small dipeptides in nonpolar solvents suggests that the methods and parameters used in calculating conformational energies are not a major problem.

(*b*) The nature of the approximations in the statistical mechanics has been treated in detail (15). It was shown (15) that a classical statistical mechanical analysis of the conformational space of *isolated* blocked dipeptides was valid using the Z-approximation (Eqs. 2 and 4) but not the *Q*-approximation (Eqs. 1 and 3). The only difference between these two approximations is the inclusion of librational entropy in the Z function. However, the relationship between the entropy of an isolated blocked dipeptide and that of a dipeptide segment in a globular protein is not clear. Therefore, $P_{Q,b}$ may be as valid as $P_{Z,b}$ for comparison with $P_{E,b}$ (15). Comparing $P_{Z,b}$, rather than $P_{Q,b}$, with $P_{E,b}$ does not change the general conclusions of this study. However, disagreement between $P_{Z,b}$ and $P_{E,b}$ is found in five dipeptides (Gly-Asp, Ala-Ala, Ser-Ala, Ser-Ser, and Asn-Ser) for which $P_{Q,b}$ and $P_{E,b}$ do agree, and agreement between $P_{Z,b}$ and $P_{E,b}$ is found in four dipeptides (Gly-Ser, Phe-Gly, Leu-Gly, and Tyr-Gly) for which $P_{Q,b}$ and $P_{E,b}$ do not agree. The result is that 25 dipeptides have values of $P_{Z,b}$ and $P_{E,b}$ that agree and 22 dipeptides have values that do not. We feel, therefore, that the major sources of discrepancy do not involve the methodology of the statistical mechanical analysis of the blocked dipeptides.

(*c*) The failure to include solvent in the calculations appears to be a major source of the discrepancies between $P_{Q,b}$ and $P_{E,b}$ in several dipeptide sequences. Of the 14 dipeptides for which $P_{E,b}$ and $P_{Q,b}$ are not in agreement and $P_{E,b} > P_{Q,b}$ (i.e., the region above the diagonal line in Fig. 1), 10 contain a highly polar group [Ser (S), Asn (N), or Asp (D)] and another three contain an aromatic group [Phe (F) or Tyr (Y)]. The largest deviations between $P_{E,b}$ and $P_{Q,b}$ in the upper left-hand region in Fig. 1 occur in Ala-Asp (AD), Ser-Asp (SD), Pro-Asn (PN), and Ala-Asn (AN). These dipeptides have relatively low calculated bend probabilities because of highly stable side-chain–backbone hydrogen bonds, which occur primarily in nonbends. If water were included, these hydrogen bonds would be destabilized, and the relative stabilities of the bends would increase. This is especially true of the Asp-containing sequences, because Asp was taken in the un-ionized form in the calculations but would be ionized in water. Solvent may also be part of the source of discrepancy between $P_{E,b}$ and $P_{Q,b}$ found in the Gly-containing dipeptides that show disagreements in Fig. 1, because Gly has no nonpolar side chain and therefore is quite polar.

(*d*) Long-range interactions undoubtedly also play an important role in many dipeptide sequences. This especially is expected to be true of Gly-containing dipeptides, because Gly has a very flat conformational energy surface (26). Interestingly, all the dipeptides below the diagonal line in Fig. 1 for which $P_{Q,b}$ is not within the statistical error of $P_{E,b}$ contain a Gly (G) residue. This suggests that long-range interactions tend to make the conformation of Gly more extended than would be predicted by the calculations. Even though Gly is well known (2–21) to occur frequently in bends, it appears that long-range interactions (and/or solvent) destabilize the bend conformations in some Gly-containing dipeptide sequences. Of course, the influence of long-range interactions is not limited to dipeptides with a Gly residue. As discussed in another paper (15), Ala-Pro (AP in Fig. 1) has values of $P_{Q,b}$ and $P_{E,b}$ that disagree, primarily because of long-range interactions.

The Dominance of Local Interactions. An important observation about Fig. 1 is that most dipeptide sequences show agreement between $P_{E,b}$ and $P_{Q,b}$. This may have significant implications in the mechanism of folding of proteins and in the stability of the native protein structure, because agreement between $P_{E,b}$ and $P_{Q,b}$ implies that bends involving these dipeptide sequences are stabilized primarily by local interactions in the proteins; i.e., in the statistical analysis, long-range interactions tend to be averaged out. Thus, bends may exist in segments of the protein chain, as a result of the local interactions, before the native structure is formed, and therefore may be involved in directing the pathway of folding (2, 3, 8, 21).

We wish to thank Drs. F. R. Maxfield, G. Némethy, D. H. Wertz, and L. G. Dunfield for helpful discussions and comments on this manuscript, and B. B. Zimmerman for assistance in preparing the manuscript. This work was supported by grants from the National Science Foundation (PCM75-08691) and the National Institute of General Medical Sciences of the National Institutes of Health (GM-14312). S.S.Z. was a National Institutes of Health Postdoctoral Research Fellow, 1974–77.

1. Venkatachalam, C. M. (1968) *Biopolymers* **6**, 1425–1436.
2. Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A. & Margoliash, E. (1971) *J. Biol. Chem.* **246**, 1511–1535.
3. Lewis, P. N., Momany, F. A. & Scheraga, H. A. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 2293–2297.

4. Kuntz, I. D. (1972) *J. Am. Chem. Soc.* **94**, 8568–8572.
5. Esipova, N. G. & Tumanyan, V. G. (1972) *Mol. Biol.* **6**, 840–850.
6. Chandrasekaran, R., Lakshminarayanan, A. V., Pandya, U. V. & Ramachandran, G. N. (1973) *Biochim. Biophys. Acta* **303**, 14–27.
7. Crawford, J. L., Lipscomb, W. N. & Schellman, C. G. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 538–542.
8. Lewis, P. N., Momany, F. A. & Scheraga, H. A. (1973) *Biochim. Biophys. Acta* **303**, 211–229.
9. Némethy, G. & Printz, M. P. (1972) *Macromolecules* **5**, 755–758.
10. Nishikawa, K., Momany, F. A. & Scheraga, H. A. (1974) *Macromolecules* **7**, 797–806.
11. Maigret, B. & Pullman, B. (1974) *Theor. Chim. Acta* **35**, 113–128.
12. Pletnev, V. Z., Kadymova, F. A. & Popov, E. M. (1974) *Biopolymers* **13**, 1085–1092.
13. Hiltner, W. A. & Walton, A. G. (1975) *J. Mol. Biol.* **92**, 567–572.
14. Howard, J. C., Ali, A., Scheraga, H. A. & Momany, F. A. (1975) *Macromolecules* **8**, 607–622.
15. Zimmerman, S. S., Shipman, L. L. & Scheraga, H. A. (1977) *J. Phys. Chem.* **81**, 614–622.
16. Zimmerman, S. S. & Scheraga, H. A. (1977) *Biopolymers* **16**, 811–843.
17. Zimmerman, S. S. & Scheraga, H. A. (1977) *Biopolymers*, in press.
18. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 222–245.
19. Burgess, A. W., Ponnuswamy, P. K. & Scheraga, H. A. (1974) *Isr. J. Chem.* **12**, 239–286.
20. Robson, B. & Pain, R. H. (1974) *Biochem. J.* **141**, 899–904.
21. Tanaka, S. & Scheraga, H. A. (1976) *Macromolecules* **9**, 812–833.
22. Ponnuswamy, P. K., Warme, P. K. & Scheraga, H. A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 830–833.
23. IUPAC-IUB Commission on Biochemical Nomenclature (1970) *J. Mol. Biol.* **52**, 1–17.
24. Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975) *J. Phys. Chem.* **79**, 2361–2381.
25. Powell, M. J. D. (1964) *Comput. J.* **7**, 155–162.
26. Zimmerman, S. S., Pottle, M. S., Némethy, G. & Scheraga, H. A. (1977) *Macromolecules* **10**, 1–9.
27. Maxfield, F. R. & Scheraga, H. A. (1976) *Biochemistry* **15**, 5138–5153.
28. Stimson, E. R., Zimmerman, S. S. & Scheraga, H. A. (1977) *Macromolecules*, in press.